

B86-148137

Applications of Molecular Connectivity
Indexes and Multivariate Analysis in
Environmental Chemistry

Minnesota Univ.-Duluth

Prepared for

Environmental Research Lab.-Duluth, MN

Jan 86



U.S. Department of Commerce
National Technical Information Service

NTIS

TECHNICAL REPORT DATA
(Please read instructions on the reverse before completing)

1. REPORT NO. EPA/600/D-86/017		3. RECIPIENT'S ACCESSION NO. PB86 148137/AS	
4. TITLE AND SUBTITLE Applications of Molecular Connectivity Indexes and Multivariate Analysis in Environmental Chemistry		5. REPORT DATE January 1986	
7. AUTHOR(S) G.J. Niemi*, R.R. Regal**, and G.D. Veith***		6. PERFORMING ORGANIZATION CODE	
9. PERFORMING ORGANIZATION NAME AND ADDRESS *Dept. Pharmacol., **Dept. Math. Sci., Univ. Minn., Duluth, MN; *** U.S. EPA, ERL-Duluth, 6201 Congdon Blvd., Duluth, MN		8. PERFORMING ORGANIZATION REPORT NO.	
12. SPONSORING AGENCY NAME AND ADDRESS Environmental Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Duluth, MN 55804		10. PROGRAM ELEMENT NO.	
		11. CONTRACT/GRANT NO. 810824	
		13. TYPE OF REPORT AND PERIOD COVERED	
		14. SPONSORING AGENCY CODE EPA-600/03	
15. SUPPLEMENTARY NOTES IN: Environmental Applications of Chemometrics, J.J. Breen and P.E. Robinson (eds.), ACS Symp. Series 292, Am. Chem. Soc., Washington, DC, pp. 148-159 (1985)			
16. ABSTRACT We have developed a data matrix of 90 variables calculated from molecular connectivity indices for 19,972 chemicals in the Toxic Substance Control Act (TSCA) inventory of industrial chemicals. Principal component analysis of this matrix revealed eight principal components that explained > 93 % of the variation in these data. The first three principal components convey generalized information on chemical structure: size, degree of branching, and number of cycles. The other components contained more specific information on branching, bonding, cyclicness, valency, and combinations of these structural attributes. Here we explored the use of the connectivity indices and their calculated principal components for their potential in predicting biodegradation as measured by biochemical oxygen demand (BOD) and the octanol/water partition coefficient. This approach showed promise in the prediction of biodegradation, but was of limited use in the prediction of the partition coefficient. Because it is possible to calculate the connectivity indices at a nominal cost for nearly all chemicals, the approach will prove especially useful for the identification of chemicals with similar structures and for systematically exploring where data are lacking on biological endpoints for chemicals in TSCA.			
17. KEY WORDS AND DOCUMENT ANALYSIS			
a. DESCRIPTORS		b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
18. DISTRIBUTION STATEMENT Release to public		19. SECURITY CLASS (This Report) Unclassified	21. NO. OF PAGES
		20. SECURITY CLASS (This page) Unclassified	22. PRICE

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ACS SYMPOSIUM SERIES 292

Environmental Applications of Chemometrics

Joseph J. Breen, EDITOR

*Office of Toxic Substances
U.S. Environmental Protection Agency*

Philip E. Robinson, EDITOR

*Office of Toxic Substances
U.S. Environmental Protection Agency*

Developed from a symposium sponsored by
the Division of Environmental Chemistry
at the 188th Meeting
of the American Chemical Society,
Philadelphia, Pennsylvania,
August 26-31, 1984



American Chemical Society, Washington, D.C. 1985

Applications of Molecular Connectivity Indexes and Multivariate Analysis in Environmental Chemistry

Gerald J. Niemi¹, Ronald R. Regal², and Gilman D. Veith³

¹Department of Pharmacology, University of Minnesota—Duluth, Duluth, MN 55812

²Department of Mathematical Sciences, University of Minnesota—Duluth,
Duluth, MN 55812

³Environmental Research Laboratory, U.S. Environmental Protection Agency, Duluth,
MN 55804

We have developed a data matrix of 90 variables calculated from molecular connectivity indices for 19,972 chemicals in the Toxic Substance Control Act (TSCA) inventory of industrial chemicals. Principal component analysis of this matrix revealed eight principal components that explained > 93 % of the variation in these data. The first three principal components convey generalized information on chemical structure: size, degree of branching, and number of cycles. The other components contained more specific information on branching, bonding, cyclicity, valency, and combinations of these structural attributes. Here we explored the use of the connectivity indices and their calculated principal components for their potential in predicting biodegradation as measured by biochemical oxygen demand (BOD) and the octanol/water partition coefficient. This approach showed promise in the prediction of biodegradation, but was of limited use in the prediction of the partition coefficient. Because it is possible to calculate the connectivity indices at a nominal cost for nearly all chemicals, the approach will prove especially useful for the identification of chemicals with similar structures and for systematically exploring where data are lacking on biological endpoints for chemicals in TSCA.

More than 50,000 chemicals are currently listed in the Toxic Substance Control Act (TSCA) inventory, but physical-chemical properties are available for a relatively small percentage and biological endpoints for even less. The costs associated with thoroughly testing all chemicals are prohibitive, so models are needed to (1) predict the environmental effects of a new chemical, or (2) assess whether the chemical should be subject to a detailed testing regime (1). Although models are available to

0097 6156/85 0292 0148\$06.00/0

© 1985 American Chemical Society

NTIS is authorized to reproduce and sell this
report. Permission for further reproduction
must be obtained from the copyright owner.

assess the environmental effects for some groups of chemicals (2-6), new compounds often cannot be categorized into one group or another. If it is unclear when a model can be applied to a new chemical, then the power of the model is substantially reduced.

To overcome this weakness, we are developing a quantitative structure-activity strategy that is conceptually applicable to all chemicals. To be applicable, at least three criteria are necessary. First, we must be able to calculate the descriptors or independent variables directly from the chemical structure and, presumably, at a reasonable cost. Second, the ability to calculate the variables should be possible for any chemical. Finally, and most importantly, the variables must be related to a parameter of interest so that the variables can be used to predict or classify the activity or behavior of the chemical (1). One important area of research is the development of new variables or descriptors that quantitatively describe the structure of a chemical. The development of these indices has progressed into the mathematical areas of graph theory and topology and a large number of potentially valuable molecular descriptors have been described (7-9). Our objective is not concerned with the development of new descriptors, but alternatively to explore the potential applications of a group of descriptors known as molecular connectivity indices (10).

Molecular connectivity indices are desirable as potential explanatory variables because they can be calculated for a nominal cost (fractions of a second by computer) and they describe fundamental relationships about chemical structure. That is, they describe how non-hydrogen atoms of a molecule are "connected". Here we are most concerned with the statistical properties of molecular connectivity indices for a large set of chemicals in TSCA and the presentation of the results of multivariate analyses using these indices as explanatory variables to understand several properties important to environmental chemists. We will focus on two properties for which we have a relatively large data base: (1) biodegradation as measured by the percentage of theoretical 5-day biochemical oxygen demand (BOD)(11), and (2) n-octanol/water partition coefficient or hereafter termed log P (12).

Data Base

The U.S. EPA Environmental Research Laboratory-Duluth with the help of their cooperators has developed a data matrix of 90 variables calculated from molecular connectivity indices (10) for 19,972 of the chemicals in TSCA. Molecular connectivity indices consist of four primary types (paths or the edges between atoms, clusters or branches, path/clusters, and cycles or rings) that are calculated from 0th to 9th order depending on the number of connections between atoms. Path terms can include as many orders as there are edges between atoms in the molecule, the minimum order for a cluster or a cycle is three, and the minimum for a path/cluster is four. Therefore, using 0th to 9th order, the number of variables for one set of connectivity indices is 30 variables. In our data base, we included three sets of

connectivity indices: simple indices in which the molecule is assumed to be a saturated hydrocarbon, bond-corrected indices that adjust for double and triple bonds between atoms, and valence-corrected indices that adjust for the heteroatoms in the molecule.

The data base for biodegradation information consisted of 5-day BOD tests that were screened from the literature using a systematic review procedure (13). A total of 340 chemicals was included in this analysis. In contrast, the data base for log P values is much larger because Leo and Weininger (14) have developed an additive model to estimate log P for chemicals. Their model also gives an estimate for the confidence in the estimated value of log P. Here we used a data base of 1700 chemicals that were determined by Leo and Weininger (14) to have a high confidence in the estimated value of log P.

Statistical Analysis

We focused on four primary objectives: (1) calculation of the statistical properties of the molecular connectivity indices, (2) evaluation of how the dimensionality (90 variables) of the molecular connectivity variables could be reduced (15-17), (3) prediction of high or low 5-day BOD using the molecular connectivity variables as discriminators (18), and (4) estimation of log P values using the molecular connectivity indices as explanatory variables (19-20).

Evaluation of the statistical properties is a fundamental part of any statistical analysis and here we concentrated on the distribution of each variable. To reduce the dimensionality of this data set we used principal component analysis (PCA) to explore the covariance structure of these data and to reduce the variables to a more manageable number (PAI method with no rotation, 21).

Because a complete explanation of the procedure used to predict the relative degree of biodegradation of chemicals is prohibitive (see 22), here we focus on the overall results. Briefly, our multivariate statistical analysis included (1) calculating the factor scores derived from six principal components of a principal component analysis extracted from 45 of the molecular connectivity variables for the 19,972 chemicals of TSCA, (2) using these six principal components to identify ten clusters in the principal component space with the K-means clustering algorithm of the Biomedical Computer Program (23), and (3) calculating the discriminant functions that best separated chemicals with high BOD values (theoretical BOD > 17%) from those with low BOD values (theoretical BOD < 13 %; there were no chemicals with BOD values between 13 and 17 %) using the molecular connectivity indices as discriminators (discriminant function analysis, 23). The BOD values were divided into high and low groups because we were most interested in the identification of whether a chemical was degradable (high BOD) or persistent (low BOD).

In comparison, log P does not follow a dichotomous logic similar to the BOD values and, therefore, we treated log P as a

continuous variable. For the predictions of log P we explored several multiple regression models (21) and used either principal components or molecular connectivity indices as explanatory variables. With the principal components we explored the use of non-linear trends by including second and third order polynomials as explanatory variables. With the molecular connectivity indices we used two different variable modifications in the regression models: (1) logarithmic transformations, and (2) standardization by size whereby the logarithm of the number of atoms was subtracted from the logarithm of each variable. The regression analyses included using both forward and stepwise methods. We were primarily concerned with finding combinations of variables that best reduced the standard error of the estimate (square root of the mean square error) (21).

Results

Statistical properties of connectivity indices. Because there are so many variables involved in this analysis, we only included the mean and standard deviation of the original and log-transformed data for the simple molecular connectivity indices (Table I). The relative magnitude of the means and standard deviations for the bond-corrected and valence-corrected indices were very similar to those for the simple indices. There are two features of these variables that have an important bearing on their statistical distribution. First, each variable is skewed to the right because of the presence of a few large molecules relative to the bulk of the chemicals in the data base. Second, many of the variables have a relatively large number of zero values. This is especially pronounced in the third and fourth order cycle terms because more than 98 % of the industrial chemicals do not have this configuration. The log-transformations improve the distributions considerably in terms of the ratio between the mean and standard deviation, but many of the variables do not approximate a normal distribution. Given these problems in the distribution of the data, we proceed with caution and consider subsequent analyses as exploratory. Furthermore, the shortcomings in the distribution of these data supports the use of PCA in the data reduction as opposed to procedures of factor analysis. Because PCA is used in an exploratory manner and no specific hypothesis is being tested (24), the assumption of normality in these data can be relaxed.

Data reduction. We used the log-transformed data in all analyses presented here. The PCA resulted in eight principal components with eigenvalues > 1 and they explained 93.5% of the variation in the original data (Table II). The first three principal components all convey generalized information on chemical structure: size (PC 1), degree of branchness (PC 2), and number of cycles (PC 3). PC 1 was positively correlated with all 90 variables ($r > .32$), except for the cyclic variables in which r was as low as .07 for the 3rd order cyclic variables. PC 2 was positively correlated ($r > .26$) with all cluster variables, but negatively correlated with all path and cyclic variables. PC 3

Table I. Statistical properties of 0th to 9th order for path, cluster, path/cluster, and cycle types of simple connectivity indices.

Term	Order	Original		Log-transformed ¹		Percentage of zeroes
		Mean	Standard deviation	Mean	Standard deviation	
Path	0	11.24	5.35	1.09	.19	0.0
	1	7.10	3.51	.91	.20	0.0
	2	6.34	3.51	.86	.21	0.4
	3	4.67	3.00	.74	.23	4.9
	4	3.45	2.51	.63	.25	10.1
	5	2.51	2.15	.52	.26	19.0
	6	1.58	1.56	.39	.25	33.3
	7	1.00	1.19	.28	.23	38.5
	8	.61	.86	.20	.20	43.5
	9	.37	.63	.14	.17	50.1
Cluster	3	1.23	1.16	.32	.18	37.6
	4	.09	.21	.03	.07	80.2
	5	.36	.93	.11	.14	51.1
	6	.09	.47	.02	.09	82.8
	7	.18	.89	.04	.12	66.5
	8	.11	.83	.02	.10	89.5
Path/ cluster	9	.13	.96	.02	.11	81.2
	4	2.35	2.68	.47	.26	18.1
	5	3.23	3.86	.54	.32	18.2
	6	4.68	6.57	.61	.40	20.2
	7	5.43	8.74	.62	.45	22.8
	8	6.09	11.80	.60	.50	25.0
Cycle	9	6.50	15.09	.56	.54	29.3
	3	.01	.04	.00	.02	98.5
	4	.01	.05	.00	.02	98.1
	5	.02	.10	.01	.03	84.5
	6	.07	.19	.05	.06	36.2
	7	.12	.36	.08	.09	40.6
	8	.17	.59	.11	.13	45.9
	9	.21	.88	.13	.17	48.2

¹ natural logarithms

Table II. Interpretations of 8 principal components calculated from 90 variables based on molecular connectivity indices for 19,972 industrial chemicals.

Principal component	Eigen-value	Variation explained %	Low values of principal component	High values of principal component
1	47.36	52.6	small molecules	large molecules
2	12.14	13.5	few branches on molecule	multi-branched molecules
3	10.53	11.7	non-cyclic molecules	multi-cyclic molecules
4	5.19	5.8	7th to 9th order cycles	3rd to 4th order cycles
5	3.13	3.5	molecules with single bonds and simple branching patterns	multi-branched molecules with double or triple bonds and/or with many heteroatoms
6	2.83	3.2	complex branching patterns and multi-cyclic molecules with few heteroatoms	complex 3rd and 4th order cyclic molecules
7	1.74	1.9	5th to 7th order cycles	complex valence-corrected branched chemicals with many heteroatoms
8	1.22	1.4	short chain molecules with complex branches	long chain molecules with few heteroatoms

was positively correlated ($r > .32$) with all cyclic variables and negatively correlated with most of the other variables. The remaining principal components explained patterns of variation among chemicals in terms of gradients of cyclicness, bonding, branching, valency (e.g., number and positions of heteroatoms such as halogens and oxygen), and combinations of these structural attributes.

The results of the principal component analysis describe the "intrinsic" dimensionality that is measured about chemical structure by the molecular connectivity indices. Eight principal components explain 93.5 % of the variation in the 90 original variables, but 23 principal components are required to explain 99.0 % of the variation in these data. The relatively high percentage of explained variation in the first three principal components (77.8 %) indicates that there are relatively high correlations among the variables derived from the molecular connectivity indices and it is obvious that size, branchness, and cyclicness are major structural properties of chemicals. In terms of data reduction, the principal component analysis has reduced the dimensionality from 90 variables to 8 new variables that still explained 93.5 % of the variation in the data set. Furthermore, if we can interpret the first three principal components as a size axis (PC 1), an axis of branchness (PC 2), and an axis of cyclicness (PC 3), then we can envision the higher order components as conveying both subtle and potentially important structural information because they are statistically uncorrelated ($r = 0.0$) or "independent" of size, branchness, and cyclicness.

Prediction of BOD value. In the ten clusters identified by the K-means clustering procedure, two clusters were represented by chemicals with only low BOD values and one cluster with nearly all (18 of 19 or 95 %) high BOD values (Table III). Therefore, no discrimination was attempted within these clusters. In the remaining clusters there were 202 high BOD chemicals and 97 low BOD chemicals. Of these, approximately 75 % (152 of 202) were correctly classified into the high BOD class, while 73 % (71 of 97) were correctly classified into the low BOD class. Using both the clustering and discrimination analyses, 77 % (170 of 220) and 78 % (93 of 120) of the chemicals in the data base were correctly classified. Within each of the clusters, between 2 and 4 molecular connectivity indices were used in the final discriminant functions to separate the two classes of BOD. Within each cluster a different combination of variables were used as discriminators. Because of the exploratory nature of this analysis, we lowered the F-ratio inclusion level to 1.0. In several of the clusters, the F-ratios for variables included in the discriminant functions were subsequently small (e.g., < 4.0).

Predictions of log P with regression. As would be expected, the largest values of the explained variation (r squared) and the smallest standard error of estimates found with the regression models were those that included all 90 variables. These models

included one with log-transformed variables and the other with both log-transformations and with each variable standardized by size (Table IV). In general, the standard errors of the estimates were large relative to the mean which indicates a relatively poor fit for the models tested here.

TABLE III. Summary of cluster analysis and discriminant function analysis of high (> 17 % of theoretical) and low (< 13 % of theoretical) biochemical oxygen demand (BOD) values for 340 chemicals.

Cluster	Cluster analysis		Discrimination analysis	
	Number of chemicals		% Correctly classified	
	High BOD	Low BOD	High BOD	Low BOD
1	0	6	-	-
2	0	16	-	-
3	5	13	60	85
4	13	9	85	78
5	18	1	-	-
6	35	18	60	78
7	10	12	80	83
8	60	15	83	53
9	29	14	72	64
10	50	16	76	75
Total	220	120	77	78

Table IV. Summary of multiple regression analyses for the prediction of log P from molecular connectivity indices (sample size = 1700).

Variable		Regression Method	Over-all F	R ² Variables		Standard Error of Estimate
Number	Modification			10	All	
90	logarithm	stepwise	34.8	.54	.62	1.26
90	logarithm	forward	30.2	.41	.62	1.26
90	logarithm+ size	stepwise	35.3	.54	.62	1.26
90	logarithm+ size	forward	30.1	.27	.62	1.26
24	polynomials-third degree	stepwise	32.1	.24	.31	1.66
8	principal component	stepwise	64.0		.23	1.75

Discussion

Of the two exploratory analyses that we presented, one showed relatively promising results for the general classification of high or low BOD values while the other with log P provided unsatisfactory results. Among the advantages of the statistical procedures we used in the BOD model may have been the effect of reducing the dimensionality (PCA) and the complexity of the problem by clustering the chemicals into smaller groups. In contrast, for the regression analyses with log P, we attempted to deal with the problem on a global scale whereby all chemicals were included in one large analysis. Suckling et al. (25) have noted this dilemma by pointing out that complex problems are difficult to solve in one piece and need to be broken down into components, especially for the purposes of modelling. In the analytical procedures of the BOD model we used this process: (1) reduced the dimensionality of the data, (2) clustered the chemicals according to similar structural properties, and (3) attempted to analyze the parameter of interest within a specific cluster or group of similar chemicals. A further advantage of this procedure is that chemicals can be assigned independently to a cluster without prior biases regarding where it "fits" within the realm of existing chemicals. This is especially important from the perspective of environmental regulation of a new chemical. A newly designed chemical by virtue of being new does not necessarily fit into one group of chemicals or another.

Principal components. One of the potential powers of the principal components is that they are calculated from a large data base representing many chemical configurations and they are independent axes (intercorrelation $r = 0.0$) (24). Therefore, they are well-suited for building prediction models with multivariate methods such as regression where a desirable property among the explanatory variables is minimal multicollinearity (26). However, in these analyses, principal components and polynomial regression of the principal components were relatively poor predictors of log P and neither appear to be useful in this application. This is a contrast with the results presented by Murray et al. (19) who found a high correlation between log P and a form of the valence-corrected molecular connectivity index. Murray et al. (19) limited their correlation analysis to distinct chemical groups such as esters and alcohols. Our analysis included all the chemicals available in a global model.

Although the inclusion of all 90 variables produced a relatively high r^2 and the lowest standard error of estimate, we suspect that there are many spurious correlations involved in the prediction equation (27). It is slightly more encouraging that the first ten variables included in the stepwise regression models of the 90 variable log-transformed and 90 variable log-and-size-transformed data sets produced r^2 values of .54. The standard error of the estimates of log P with the regression equations is substantially higher than the estimates obtained by Leo and Weininger (14) in their model.

Exploratory data analysis and data feedback. There are a large number of variables that can be calculated from a chemical structure, yet when the number of variables becomes large relative to the sample size then spurious or trivial correlations increase (27). For example, we believe the multiple regression analyses of log P with 90 variables may represent such a situation. Likewise, the DFA used in the BOD model examined a large number of variables relative to the sample size within some of the clusters (e.g., see cluster 3, Table III). In actuality, we used the DFA as an exploratory tool to find potential molecular configurations that were associated with high or low BOD values (22). This exploration led to the identification of several subgraphs of a molecule that may be associated with persistence or degradability of a chemical. The next step in the process would be to identify chemicals with these specific subgraphs and test them for their relative degradability.

We believe that two of the main limitations for progress in applications of multivariate analyses in environmental chemistry at the current time are (1) the lack of a large data base (hundreds or thousands of chemicals) that has been collected systematically for endpoints of environmental concern (e.g., biodegradation, toxicity, and carcinogenicity), and (2) the lack of a feedback mechanism whereby new chemicals or chemicals with specific molecular configurations are being retested for specific biological endpoints. Regarding the former, we recognize that there are many limitations in the BOD data base (e.g., 28); however, there are no alternative data bases with a sample size in the hundreds that can be used in attempts to model the important environmental process of biodegradation.

Future directions

We are pursuing analyses with a variety of additional endpoints, but it is already clear that connectivity indices and multivariate techniques will be useful in some applications and limited in others. A powerful practical application of this approach is that many variables can now be calculated for nearly all chemicals. Therefore, the "universe" of industrial chemicals in TSCA can be defined in multi-dimensional space. This space gives EPA a tool for the identification of (1) whether there is a similarly-structured chemical to a new chemical that needs to be evaluated, or (2) where data on biological endpoints are lacking within this chemical space. The former concept of similarly-structured or analogous chemicals having similar biological activity is at the heart of the structure-activity perspective for predicting the environmental effects of chemicals. The latter provides an effective, objective means for the identification of what chemicals need to be tested. Both tools will be useful in understanding the relationships between chemical structure and biological activity and in applications for the assessment of the environmental effects of chemicals. One of the most important aspects, however, will be to understand the capabilities of the approach and especially its limitations.

Kowalski (29) has correctly emphasized that "the measurements made by analytical chemists are associated with some degree of uncertainty," so "it is difficult to conceive of a more perfect marriage than analytical chemistry and statistics and applied mathematics." The marriage is the study of chemometrics and we are likely "only at the threshold of realizing the importance of statistical and mathematical techniques" (30) in applications of chemical measurements. Multitudes of data are currently capable of being generated by sophisticated analytical equipment. Similarly, adaptable software and high-speed computers are able to handle these large data sets. The next decade promises to be the "decade when chemistry advances as a multivariate science (31)."

Acknowledgments

This research was supported by a cooperative agreement (CR810824-0190) between the U.S. Environmental Protection Agency and the University of Minnesota, Duluth.

Literature Cited

1. Veith, G.D. "State-of-the-Art Report on Structure-activity Methods Development"; Environmental Research Laboratory, Duluth, 1980.
2. Ogino, A.; Matsumura, S.; Fujita, T. *J. Med. Chem.* 1980, 23, 437.
3. Yuan, M.; Jurs, P. *Toxicol. and Appl. Pharmacol.* 1980, 52, 294.
4. Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J.J. *Ind. J. Chem.* 1981, 20B, 894.
5. Birge, W. J.; Cassidy, R. A. *Fund. and Appl. Toxicol.* 1983, 3, 359.
6. Gillett, J. W. *Environ. Toxicol. and Chem.* 1983, 2, 463.
7. Balaban, A. T. *Theoret. Chim. Acta.* 1979, 53, 355.
8. Sabljic, A.; Trinajstić, N. *Acta Pharm. Jugosl.* 1981, 31, 199.
9. Bertz, S. H. "A Mathematical Model of Molecular Complexity"; King, R. B., Ed.; *STUDIES IN PHYSICAL AND THEORETICAL CHEMISTRY* Vol. 28, Elsevier Science Publishers, Amsterdam, 1983; pp. 206-221.
10. Kier, L. B.; Hall, L. H. "Molecular Connectivity in Chemistry and Drug Research"; Academic Press, New York, 1976, p. 257.
11. Sawyer, C. N.; Bradney, L. *Sew. Works J.* 1946, 18, 1113.
12. Hansch, C.; Quinlan, J. E.; Lawrence, G. L. *J. Org. Chem.* 1968, 33, 347.
13. Vaischnav, D. "Biochemical oxygen demand data base"; Call, D.J.; Brooke, L.T.; Vaischnav, D.; *AQUATIC POLLUTANT HAZARD ASSESSMENT AND DEVELOPMENT OF HAZARD PREDICTION TECHNOLOGY BY QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS*. University of Wisconsin, Superior research project report (CR809234) 1984.
14. Leo, A.; Weininger, D. "CLOGP Version 3.2 User Reference

- Manual"; Medicinal Chemistry Project, Pomona College, Claremont, 1984.
15. Cramer, R. D. *J. Am. Chem. Soc.* 1980, 102, 1837.
 16. Lukovits, I.; Lopata, A. *J. Med. Chem.* 1980, 23, 449.
 17. Burkhard, L. P.; Andren, A. W.; Armstrong, D. E. *Chemosphere* 1983, 12, 935.
 18. Geating, J. "Project Summary: Literature Study of the Biodegradability of Chemicals in Water: Vols. 1 and 2"; U.S. Environmental Protection Agency, Cincinnati, EPA-600/s2-81-175/176, 1981, p. 4.
 19. Murray, W. J.; Hall, L. H.; Kier, L. B. *J. Pharm. Sci.* 1979, 64, 1978.
 20. Govers, H.; Ruepert, C.; Aiking, H. *Chemosphere* 1984, 13, 227.
 21. Nie, N. H.; Hull, C. H.; Jenkins, J. G.; Steinbrenner, K.; Bent, D. H. "Statistical Package for the Social Sciences"; McGraw-Hill, New York, 1975, p. 675.
 22. Niemi, G. J.; Regal, R. "Predicting Biodegradability from Chemical Structure: the Use of Multivariate Analysis"; U.S. Environmental Protection Agency, Duluth, 1983.
 23. Dixon, W. J.; Brown, M. B. "Biomedical Computer Programs - P Series"; University of California Press, Berkeley, 1979, p. 880.
 24. Morrison, D. F. "Multivariate Statistical Methods"; McGraw-Hill, New York, 1967, p. 338.
 25. Suckling, C. J.; Suckling, K. E.; Suckling, C. W. "Chemistry Through Models"; Cambridge University Press, Cambridge, 1978.
 26. Tatsuoka, M. "Multivariate Analysis"; J. Wiley, New York, 1971.
 27. Wold, S.; Dunn, W. J. III. *J. Chem. Inf. Comput. Sci.* 1983, 23, 6.
 28. Howard, P. H.; Saxena, J.; Sikka, H. *Environ. Sci. Technol.* 1978, 12, 398.
 29. Kowalski, B. R. *Anal. Chem.* 1980, 52, 112R.
 30. Kowalski, B. R. *Anal. Chem.* 1978, 50, 1309A.
 31. Kowalski, B. R.; Wold, S. In "Handbook of Statistics, Vol. 2"; Krishnaiah, P. R.; Kanal, L. N. Eds.; North-Holland Publishing Co., 1982; Chap. 31.

RECEIVED June 28, 1985

Reprinted from ACS SYMPOSIUM SERIES No. 292
Environmental Applications of Chemometrics
Joseph J. Breen and Philip E. Robinson, Editors
Copyright © 1985 by the American Chemical Society
Reprinted by permission of the copyright owner

END

DATE

FILMED

3-27-86

NTIS